BEHAVIORAL AND SUBJECTIVE WORKLOAD METRICS FOR OPERATIONAL ENVIRONMENTS

by
Clark A. Shingledecker, Ph.D.
Air Force Aerospace Medical Research Laboratory
Workload and Ergonomics Branch, Human Engineering Division
Wright-Patterson AFB, Ohio 45433

SUMMARY

The assessment of crew performance capability under conditions of sustained intensive air operations requires the use of specialized measures of operator workload which are matched to the nature of the investigation and to the environment in which the workload evaluation must be conducted. In many cases, the effects of severe combined stressors and of aircrew performance requirements on mental workload cannot be studied in the laboratory, and must be addressed in high fidelity simulation or during operational test exercises. This paper examines the advantages and limitations of traditional subjective report and behavioral measures of workload for application in operational environments. In addition, recent efforts at the U.S. Air Force Aerospace Medical Research Laboratory to develop improved field-usable subjective and behavioral secondary task metrics are described.

INTRODUCTION

The management of mental workload is of central importance to the successful design and deployment of modern manned airborne weapon systems. Although the maintenance of an effective military force requires the development of systems which incorporate the most sophisticated products available from the engineering disciplines, such technical improvement is often achieved with a simultaneous growth in the monitoring, supervisory and decision making responsibilities of the aircrew. These information processing requirements can jeopardize system performance by placing demands on the human operator which approach, or even exceed, his finite mental capacities. To compound the problems induced by complex system technology, workload may also be affected by the projected nature of the battle environment in which military aircrews are expected to function. Even when it can be demonstrated that an aircraft system is operable under relatively benign conditions, the stresses produced by wartime operations may catastrophically increase the likelihood of failures in human performance. Such intensive battle environments threaten to compromise aircrew capabilities in at least two ways. Stressors such as fatigue, sleep loss, personal threat, and chemical, biological or nuclear agents may reduce the perceptual, cognitive, and motor capacity of the individual. At the same time, the dynamic nature of modern combat and the rapidly changing tactical situation will increase uncertainty and add to the operator's already high mental processing load.

The achievement of mission effectiveness under the circumstances outlined above requires careful consideration of aircrew capability during all phases of the aircraft development process and throughout the operational life of the system. More specifically, the selection of appropriate crew station design features or the optimization of deployment strategies depends upon the availability of accurate and reliable methods to assess mental workload in a variety of evaluation environments. Unfortunately, no widely accepted standards for workload measurement presently exist, and the choice of a workload metric is often left to the discretion and experience of the individual scientist or engineer assigned to a workload evaluation problem.

Workload science provides two major thrust areas which are contributing to the design of a coherent workload assessment technology. The first involves research directed toward the development of a workload theory or model which would accurately characterize the functional operation of the human information processing system, and describe the nature of the capacities of this system which limit performance. Although no universal consensus has been reached on these theoretical issues, recent evidence offers strong support for a multiple capacity model (1, 2). A second line of workload research is concerned with the development and application of specific measurement techniques. These metrics can be classified by the type of human response used to make inferences about workload, and include indices from the behavioral, physiological and subjective domains. While guided to some extent by theoretical trends, workload metrics have been selected and developed largely through individual experiments demonstrating a correspondence between the measured index and some manipulation of the parameters of one or more tasks that were hypothesized to influence processing demands. Application of this general methodology has led to the identification of a variety of techniques purported to assess workload, and recent reviews have provided an extensive catalogue of these metrics (3).

SELECTING WORKLOAD MEASURES

Although a completely adequate model and a standardized methodology await future scientific developments, several guidelines for the use of workload measures can be derived from our current theoretical understanding of mental capacity and from a knowledge of the practical constraints of applied research. Such considerations indicate that no single measure is likely to be sufficient for every evaluation environment or for addressing all workload questions. From a functional standpoint, little empirical support exists for a belief that one sort of metric will produce a "better" estimate of workload than another. The ultimate purpose of any workload measure is to predict performance failure, and an index which serves this function may be equally likely to be found in a physiological parameter, a behavioral response, or a subjective report. However, these general types of measure, as well as specific varieties within each class, have the potential for differing along several important dimensions. Listed below are seven such dimensions which are relevant to the selection of an appropriate workload measure.

Validity
Reliability
Sensitivity
Diagnosticity
Intrusiveness
Acceptability
Instrumentation Requirements

The simplest features that any metric should possess are the classical properties of validity, reliability and sensitivity. However, due to the multidimensionality of the phenomenon of workload, these standards alone are insufficient. As noted earlier in this paper, although cognitive science has not produced a universally accepted model of mental resources, recent research tends to support a multiple capacity view of workload. This model holds that performance limitations can arise from insufficient mental resource availability in any of a number of independent processes. The implication for workload assessment is that different tasks performed by the operator will generate variable patterns of load with respect to each resource. As a result, it is likely that workload measures will vary in their diagnosticity (2), or the degree to which they discriminate among workload on individual resources. In reality, this factor can be conceived of as a broad dimension with some metrics providing indices of specific resource demands and others offering more global measures of workload.

In addition to these theoretical properties, workload metrics also vary along pragmatic dimensions. Regardless of the validity or sensitivity that a measure might possess, its ultimate value depends on whether it can be usefully employed in the particular situation in which a workload question must be addressed. Workload metrics differ widely in their intrusiveness or propensity to interfere with ongoing primary task behavior. While intrusion is of minor importance in workload research conducted under controlled laboratory conditions, any sacrifice of performance efficiency may be intolerable in potentially dangerous system test conditions or when primary task performance measures are of interest. Metrics also vary in the degree of operator acceptance that they enjoy and in the amount of instrumentation required to obtain them. Again, while neither of these factors pose problems in a laboratory environment, the introduction of a workload measure which produces discomfort or distraction may prove to be of little worth under field test conditions. Likewise, a measure which requires bulky or fragile equipment will not be usable in situations such as flight test.

It should be noted that in some cases these dimensions are not independent. One possibility is that sensitivity and diagnosticity are inversely related so that highly diagnostic measures are likely to be more sensitive to workload than those which provide a more general index. Taken together however, the classification criteria offer a practical method for comparing workload metrics, and form a basis for selecting a measure to meet the needs of specific workload assessment problems. The value of characterizing metrics along these dimensions can be appreciated by considering the trade-offs among metric qualities that might be required to address the range of workload questions that arise in airborne weapon systems. For example, an easily obtained, nonintrusive, simply instrumented metric having the capability to assess workload across a large class of tasks with a moderate degree of sensitivity might be needed to isolate the high workload segments of a full mission operational system test. At the other extreme, a completely different set of more complex and restrictive metrics with the capabilities to identify mental resource loading patterns and to make fine discriminations among workload levels is more appropriate in a laboratory or part-task simulation environment where the purpose is to isolate causal factors and to select design options which alleviate workload.

Solutions to the problem of maintaining human performance capability under the conditions of sustained intensive air operations will require the use of a number of measurement techniques which can be employed to answer workload questions in a broad range of testing environments. In some cases, these questions can be addressed in the laboratory. For example the heat stress produced by chemical defense ensembles might be expected to impair the information processing capacity of the operator, thereby increasing the workload of a flight control task. While the operator may be able to compensate and maintain control performance under such stress, it is necessary to determine the degree to which various ensemble designs limit capacity in order to predict his ability to engage in other duties or to deal with unexpected mission events. Such well-defined problems can be fruitfully studied under controlled part-task conditions, and the scientist has a wide selection of workload metrics from which to choose that vary in diagnosticity and technical complexity.

However, a number of crucial human performance issues surrounding sustained operations are unamenable to basic investigation. Meaningful answers to questions such as the maximum number of missions a crew member should be expected to fly over a brief period of time depend upon a host of interacting factors related to the man-machine interface and to the multiple physiological and psychological stresses present in an actual operational environment. In these cases valid measurements must be obtained in high fidelity simulation studies or during operational military exercises. With respect to the classification scheme described earlier, these testing conditions require metrics which are nonintrusive, highly acceptable to operators, and which necessitate the use of minimal experimental equipment to collect workload estimates. Unfortunately, as noted by Schiflett (4), most available workload measures have been developed for, and are most applicable to the laboratory environment, and are not designed to provide these qualities.

In an effort to provide answers to workload questions which arise in operational environments, the U.S. Air Force Aerospace Medical Research Laboratory has begun the development and evaluation of two workload measurement technqiues which are particularly suitable for use in high fidelity simulation or field test conditions. The first of these is a workload rating method known as the Subjective Workload Assessment Technique (SWAT) (5, 6).

SUBJECTIVE MEASURES

Subjective measures, derived from the analysis of verbal reports or from more structured rating scales, are the oldest and most widely used methods of assessing workload. The popularity of subjective metrics is attributable to their inherent advantages over other approaches. Foremost among these is ease of administration. Subjective metrics require a minimum of experimental apparatus and can be obtained under almost any conditions with little or no intrusion on primary task performance. Furthermore, these metrics take advantage of the ability of humans to integrate perceptions over flexible periods of time. As a result, subjective metrics can be focused to assess the workload of a single brief event or of a prolonged segment of task activity. Finally, subjective metrics are potentially applicable across a wide-range of operator tasks and assessment situations.

Unfortunately, while subjective metrics offer a number of ideal characteristics for workload estimation in operational environments, they also suffer from several deficiencies which limit their utility. A general problem with these measures has been a lack of standardization. Possibly because of the ease with which subjective metrics can be constructed, the workload literature contains almost as many different scales and checklists as there are subjective workload studies. This state of affairs makes it difficult or impossible to compare data collected in different experiments. A second problem with traditional subjective measures is the dilemma that arises when both sensitivity and usability are considered in the construction of a scale. If a subjective scale is designed to permit the user to make fine discrimations among levels of perceived load, it will tend to include a large number of scale points and will be liberally annotated with descriptors to insure reliability. However, such scales are often awkward to apply and require a large amount of time to collect a single rating. A final and more serious drawback of subjective approaches for assessing workload is the low level of quantification that they provide. Typically, subjective scales permit only an ordinal level of measurement to be achieved. That is, ratings can be interpreted as higher or lower than one another, but no rigorous determination can be made of the magnitude of any difference in ratings. This problem restricts the range of statistical treatment that can be applied to the data and seriously limits the type of conclusion that can be drawn from an experiment.

SWAT was specifically designed as a candidate standardized metric which would overcome many of the problems traditionally associated with subjective approaches. The technique is based on a current conceptualization (7) in which the perception of workload arises from three major factors: Time Load or the degree of temporal compression and overlap among tasks; Mental Effort or the perceived complexity and difficulty of the information processing required to perform an activity; and Psychological Stress or the combined emotional and physical factors which can otherwise affect the operator's performance capacity. In SWAT these factors are represented individually on the three-part rating scale presented below.

I. Time Load

1. Often have spare time. Interruptions or overlap among activities occur infrequently or not at all.

2. Occasionally have spare time. Interruptions or overlap among activities occur frequently.

3. Almost never have spare time. Interruptions or overlap among activities are very frequent, or occur all the time.

II. Mental Effort

1. Very little conscious mental effort or concentration required. Activity is almost automatic, requiring little or no attention.

2. Moderate conscious mental effort or concentration required. Complexity of activity is moderately high due to uncertainty, unpredictability, or unfamiliarity. Considerable attention required.

3. Extensive mental effort and concentration are necessary. Very complex activity requiring total attention.

III. Psychological Stress

1. Little confusion, risk, frustration or anxiety exists and can be easily accommodated.

2. Moderate stress due to confusion, frustration or anxiety. Noticeably adds to workload. Significant compensation is required to maintain adequate performance.

3. High to very intense stress due to confusion, frustration or anxiety. High to extreme determination and self-control required.

An immediately apparent feature of the SWAT rating scale is the ease with which relatively detailed measures can be obtained. Considering all possible combinations of ratings on time, effort and stress, the scale yields 27 discriminable points. Nevertheless, the operator can select any of these points merely by making simple choices from the three levels of each dimension. The primary advantage of SWAT however, is the multidimensional scaling method that is used to transform ordinal ratings on the three dimensions to single values on an overall interval level scale of workload. This scaling tehnique is based on a type of fundamental measurement known as simultaneous conjoint measurement (8). In order to develop a unitary scale using conjoint measurement, an expression of the internal model used by operators to combine the three dimensions is obtained from a ranking of the 27 combinations. These data are then tested against a series of mathematical axioms to establish the combinatorial rule which governed the original orderir . Several possible composition rules can be specified using this procedure. However, a simple additi.. rule has proven to be sufficient to describe the data obtained in applications of SWAT.

Once the combinatory rule has been determined, conjoint scaling procedures are applied to assign numerical values to each of the combinations which fit the rule, preserve the original ordering of the data, and insure interval scale properties.

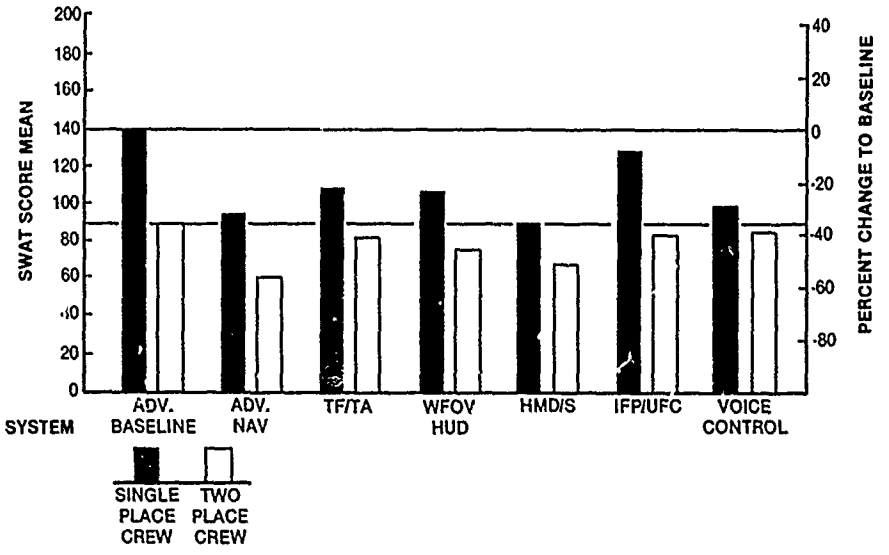
In practice, the application of SWAT to workload assessment in simulator or field tests proceeds in two stages. Prior to using the scales for event scoring, operators are asked to rank the 27 combinations of time, effort, and stress levels along a global dimension of workload. This is accomplished using 27 separate note cards on which the scale combinations are printed. In addition to providing input for the conjoint measurement and scaling algorithms, this one-hour procedure serves to familiarize the operators with the rating scales and verbal descriptors. A flexible feature of this scale development phase is that, depending upon the agreement among rankings, an overall numerical scale can be developed or separate scales can be produced for subgroups or individual operators. In a typical second stage of applying SWAT, the ordinal rating scales are employed by operators following a period of task performance to assign values of one, two, or three to each of the dimensions of time effort and stress. These ratings are then transformed to a single interval level numerical value by referring to the derived scale generated during the scale development phase.

Initial evaluation studies of the SWAT methodology have shown it to be sensitive to the workload generated by a variety of laboratory experimental tasks including both a manual control task (9) and a short term memory task (10). In addition, active duty USAF fighter and refueling aircraft pilots have participated in scale development exercises and have used the rating scales to provide workload estimates in simulation studies. In all cases, the SWAT procedures were highly acceptable to the operators and they were able to use the rating scale with little or no difficulty. Specific system problems addressed with SWAT have ranged from a comparison of the workload induced by the use of different missile systems for air-to-air combat, to a unique application in which SWAT was used in a projective fashion to predict the workload associated with various system modifications (11).

The latter study was performed as part of a USAF program designed to explore methods of enchancing future fighter aircraft systems using improved sensors, avionics and crew station design features. Once a set of enhancement options had been derived, a formal methodology was developed to comparatively assess their tactical worth, cost realism, and crew compatibility in order to guide the investment of resources on the development of prototypes and on simulation. SWAT was selected to predict the effects of each of these options on pilot workload. Nine experienced USAF pilots from an operational fighter wing completed the SWAT scale development phase and were then presented with an air-to-ground attack mission scenario and a description of a current baseline aircraft system and an advanced verison of the baseline system. The pilots were asked to provide SWAT ratings for the mission utilizing the baseline systems with one and two crew members, and for the advanced baseline system utilizing each of six enhancement options: an advanced navigation sytem (ADV NAV), an automated terrain following/terrain avoidance system (TF/TA), a wide field of view head up display (WFOV HUD), a helmet mounted display/-sight (HMD/S), and up-front control panel (IFB/UFC), and a voice actuated control system. These enhancements and their potential effects on the mission were presented to the pilots using verbal briefings and graphic materials representing displays, controls, mission profiles, and the functional operation of each option. The SWAT workload scores obtained in this study are shown in Figure 1.

FIGURE 1

AIRCREW WORKLOAD RESULTS:
NON-SENSOR ENHANCEMENTS



Since the SWAT scores were expressed as interval level numerical values, they were usable in a multi-attribute utility analysis along with predicted cost and system performance model results to permit quantitative screening of the options and to select an optimal subset for further exploratory research.

In addition to a variety of operational applications such as that described above, current developmental work with SWAT includes an effort to increase the number of rating points on each of the dimensions without sacrificing the ease with which the scale development and event scoring phases are accomplished. Furthermore, an attempt has been made to improve the sensitivity of the scale development process through the design of a method for characterizing operators in terms of six prototypical weighting strategies on the time, effort, and stress dimensions.

BEHAVIORAL MEASURES

Although appropriately designed subjective measures offer many desirable features for workload assessment in operational environments, several arguments can be made for the employment of workload metrics in these situations which are based on an evaluation of some aspect of an operator's overt task behavior. Typically, a behavioral metric is expressed in terms of some quantitative index of the speed or accuracy of performance. Because of the objective nature of such data, behavioral metrics are not likely to be influenced by extraneous motivational factors such as preference or by misinterpretation on the part of the subject or the experimenter. In addition to the relative immunity of these indices from contaminating factors, the fact that behavioral measures are derived from human performance provides a rational argument for their use in workload assessment. Since the goal of any workload index is to predict performance degradation, a metric which is based on performance might be expected to provide more useful insights to the phenomenon than other less direct methods.

Possibly the strongest support for the use of behavioral workload metrics comes from recent studies which have revealed considerable dissociation among workload measures (2). These data indicate that physiological, subjective, and behavioral metrics generally differ in the type of information that they provide. While most physiological and subjective measures appear to be general scalar indices of loading and may be insensitive to demands made on specific mental resources, behavioral measures tend to sacrifice global sensitivity for greater diagnostic capability. Thus, if a workload question demands the prediction of residual performance capacity on a number of different tasks, a behavioral measure may offer more meaningful results.

A well-known conceptual framework upon which behavioral workload measures are based portrays the human operator as a limited capacity information processing device. According to this general model, workload may be defined as the degree to which the operator's capacity is occupied by mental activities. Overload, and resulting performance decrement, occurs when capacity is insufficient to meet task demands. Since the momentary capacity of the operator is unknown and submaximal workload levels cannot be inferred from his or her performance on the primary task of interest, an indirect measure can be obtained by evaluating the amount of spare capacity available under a given set of task conditions. Although the general concept of spare capacity was derived from an early single channel model of the information processing system (12), this notion can also accommodate more complex models which propose multiple resources (2).

The behavioral approach to assessing spare capacity involves the use of the secondary task technique. In this method, operators are given an additional information processing task to perform in conjunction with the task of interest. The rationale underlying the use of secondary tasks is that by applying an extra load which produces a total information processing demand that exceeds the operator's capacity, workload can be measured by observing the difference between single and dual task performances. Normally, when secondary tasks are used as workload measures, performance on the primary task is emphasized and secondary task performance is observed as an index of the workload of the primary task.

Although the secondary task technique offers an objective and diagnostic method of assessing workload, three specific problems are encountered when traditional laboratory secondary tasks are considered for use in high fidelity simulation or operational environments. One practical problem is the physical instrumentation of the secondary task. In a flight environment, and to a lesser extent in a simulator, introducing or adding any extra equipment to a crew station may be unacceptable. The space required for electronic data recording and experimental control equipment as well as display and input devices may not be available in an already crowded cockpit. Even when sufficient space can be reserved, the possibility of obstruction or distraction caused by the additonal instrumentation can limit the feasibility of using a secondary task.

A second problem with the implementation of secondary tasks is the possibility of intrusion on primary flight duties. As noted by Gopher and North (13) primary task decrement appears to be the rule rather than the exception with the secondary task paradigm. Although some minimal level of loss in performance quality might be tolerable in an operational environment, task interference can easily complicate the interpretation of test data in situations where measures of all performance variables may be unavailable. Of course, a more serious consequence of primary task intrusion in a flight test environment is the potential for compromising flight safety. A related factor which limits the use of secondary task measures is operator acceptance. As noted by Ogden, et al (14), a secondary task is likely to produce misleading data if the operator fails to integrate it with his normal duties. Acceptance is a potential problem with all common laboratory tasks because they are obvious artificial additions to the crew station and have little face validity or congruence with the general performance situation. Such test conditions might lead the operator to neglect the secondary task or, because of its novelty, allow it to assume an artificially high priority. Thus, lack of operator acceptance can become a major contributor to primary task intrusion as well as a source of measurement error.

An analysis of the problems associated with the practical use of traditional laboratory secondary tasks prompted the development of a program of research at the U.S. Air Force Aerospace Medical Research Laboratory to explore the feasibility of designing an embedded secondary task methodology for simulation and flight test environments. The concept of the embedded secondary task is based on the hypothesis

that instrumentation limitations, task intrusion, and poor operator acceptance can be minimized by designing secondary tasks which are fully integrated with system hardware and with the crew member's conception of his mission environment. By their nature, such tasks would be realistic components of crew station activity, yet their performance could be manipulated and measured independently of the primary activities of interest.

Several classes of aircrew activity such as stores management or threat monitoring are potential candidates for isolation and use as embedded tasks. The crew station behavior that was chosen for an initial evaluation of the embedded task concept was radio communication. The communication activities which were expected to be most useful for this purpose are those initiated by a message sent from another aircraft or a ground controller to a pilot whose workload is to be assessed. Upon detection and identification of a relevant message, the pilot must engage in a sequence of verbal responses and radio switching activities in order to meet the demands of the communicated request. Such tasks closely resemble the nonadaptive discrete secondary tasks used in numerous workload studies and have many properties of good measurement tasks. Communications call upon a wide variety of information processing abilities and can be varied along several dimensions of complexity. Furthermore, no auxiliary crew station equipment is necessary to control the experiment or to collect performance data. The opportunity for obstruction or peripheral interference is also minimized since the auditory channel is not shared by other tasks and verbal responses are generally unique to radio communications activities, while switch actions can be dealt with by the pilot's free hand. Most importantly, communications tasks are an integral part of a pilot's inflight duties. As a result, lengthy training requirements are eliminated and high face validity is achieved. Additionally, the realistic nature of the activity makes artifical task interactions improbable because the pilot has predetermined priorities assigned to communications and other cockpit functions. This feature makes communications activities especially suitable for use as secondary tasks since pilots consider them to be important, but will normally devote less attention to communications as more crucial tasks become difficult to perform.

In a preliminary evaluation of realistic communications for use as embedded secondary tasks, Shingledecker, et al (15) interviewed operational A-10 pilots to obtain sample radio communications tasks from a typical air-to-ground attack misison. Each task was specified in terms of an input message and the detailed verbal and manual responses required of the pilot. Thirteen tasks associated with identification friend or foe (IFF) demand, threat alert, traffic control, waypoint passage, jammed communications, and strike clearance were selected for analysis. Examination of these tasks revealed a problem with the embedded secondary task technique which does not exist in standard laboratory tasks. While the use of realistic tasks offers many advantages, this quality also makes the tasks unamenable to precise experimental control of task demand. Traditional secondary tasks are designed to impose constrained and easily describable requirements on the performer. Thus, task parameters are easily varied and can be selected to permit precise control of loading. In contrast, realistic communications are complex processing tasks which vary along multiple dimensions. As a result, no obvious scheme can be employed to choose sets of tasks with equivalent task demand characteristics. Furthermore, excessive use of repeated task presentations must be precluded since this method of controlling task demand could sacrifice face validity. In order to resolve this dilemma, an attempt was made to scale the workload of the A-10 communications tasks. The purpose of the effort was to derive estimates of the loading associated with each task so that they could be combined in a realistic scenario in order to produce controlled levels of subsidiary task demand. Since no single a priori approach to workload evaluation was expected to produce a superior quantitative estimate, three techniques were used to provide alternative measures for later comparison to performance data.

Because workload associated with communications tasks was assumed to be partially determined by task information transmission requirements, the first scaling approach was based on information theoretical measures. The tasks were analyzed by subdivision into activities requiring perceptual decisions and those requiring manual and verbal action decisions. Each decision was then reduced to a bit measure under strict assumptions of equiprobability of alternatives and independence of sequential actions. Information transmission demands were then calculated for each task to obtain scale values. A second scaling technique was used to generate more comprehensive estimate of loading by deriving weights for information processing activities not accounted for in the first effort, and adding them to the information theoretical scale values. In order to derive weights for the demands of information gathering activities, memory requirements, and instructional complexity, 15 messages which varied along these dimensions were extracted from the sample tasks and arranged in a paired-comparisons format. Forty A-7 and A-10 pilots were asked to examine each of the 105 pairs and to indicate which of the the two entailed the greater workload. An interval scale was derived from the data using Thurstone's Law of Comparative Judgment. The scale values for the 15 messages were then used to produce weights for the extra processing activities by generating a set of simultaneous equations where the summed effects of each activity were set equal to the total scale value. A second hybrid scale was derived by selectively adding the weighting factors to the normalized bit scores for each of the original tasks. The final scaling approach tapped the subjective component of workload. Thirty pilots were asked to examine each of the 13 complete communications tasks and to rank them according to workload. An extension of Thurstone's technique was used to derive the third set of a priori estimates. The three scaling techniques were found to generate fairly consistent estimates of the workload produced by the communications tasks. Kendall's coefficient of concordance revealed a significant level of agreement among the information theoretical, hybrid, and subjective scale values ($W = .929$, $p. < .01$).

The results of the scaling efforts which showed high internal consistency among a priori estimates of the workload associated with communications tasks indicated that it was possible to construct or select tasks with controlled levels of task demand. In order to determine whether the A-10 communications tasks could be successfully embedded into an aircrew task complex and used to produce sensitive measures of workload, an experiment was conducted in an austere simulation environment. In this study subjects were required to perform eight communications tasks simultaneously with a rudimentary flight control task. The communications panels from the A-10 aircraft were installed in a fixed-base cockpit along with a joystick and a cathode ray display for presentation of the control task. The panels were wired for recording

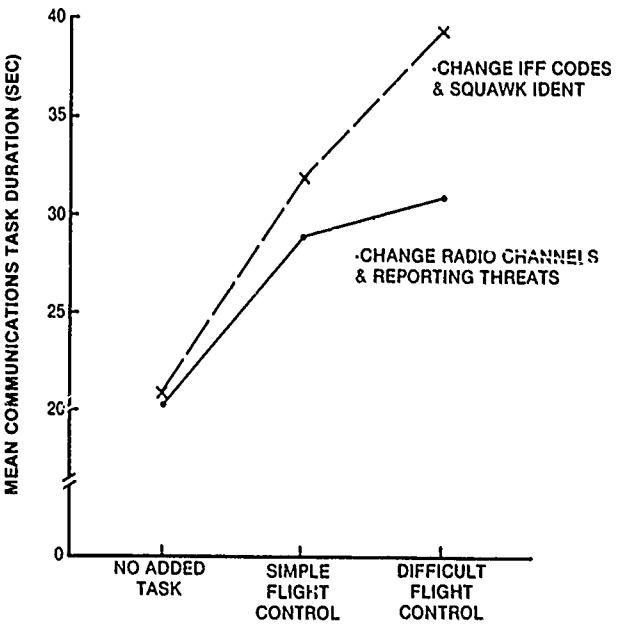
and timing of switch actions, and a communiations system was constructed to permit presentation of radio calls and recording of a subject's verbal responses.

Six paid civilian subjects participated in the study. Following familiarization with the cockpit displays and controls and with the communications tasks, each subject received approximately three hours of single task practice on the control and communications tasks. A final hour of training was devoted to dual task practice. The primary control task was a single axis compensatory task in which the stability of the controlled element could be varied to manipulate demand. During testing the subjects performed each of the eight communications tasks both singly and in combination with the control task at a low and a high level of instability (60% and 95% of a subject's maximum control capacity). Several specialized time and accuracy measures were collected for each communications task. However, in order to permit meaningful comparisons among tasks, a common measure of total task performance time was used as the predicted index for workload assessment.

Multivariate statistical analyses of the data revealed significant effects of both control task demand and of the type of communications task performed ($p < .01$). Planned comparisons performed on the differences between single task performance and dual task performance with the two levels of control task difficulty indicated that four of the eight communications tasks provided sensitive indices of workload. In addition, for two of these tasks, the control task loading effects, were displayed primarily in the performance of the embedded secondary communications task (see Figure 2).

FIGURE 2

COMMUNICATION TASK PERFORMANCE AS A MEASURE OF COCKPIT WORKLOAD



The agreement between communications task performance and the workload estimates computed in the previous scaling effort was also assessed. A significant correlation was obtained with the information theoretical scale ($r = .64$, $p < .05$) indicating that this method would be useful in constructing additional tasks.

The results of the study described above indicate that embedded radio communications tasks represent a viable approach to workload assessment in complex simulation and operational test environments. Further research is required in a high fidelity simulation setting using operational pilots as subjects in order to insure that these tasks do not produce significant interference with primary flight duties. A favorable outcome of such a study would permit the development of a general methodology for constructing embedded communications tasks which could be used by field level human factors specialists to create workload assessment techniques tailored to specific aircraft, missions, and crew stations.

CONCLUSIONS

Like other military activities, the achievement of mission effectiveness during sustained intensive air operations is dependent on the optimization of a number of crucial system and tactical parameters. However, because of its unique combination of physiological and psychological stresses, this combat

environment creates an especially sensitive link between the performance of the human operator and mission success. To maximize the operator's capability under such extreme conditions, intelligent decisions must be made with respect to crew station design variables and personnel deployment strategies. In some cases, appropriate models of human performance and available data bases on the sensory, perceptual, cognitive and motor capacities of the human operator can be used to guide these decisions. In many other situations, these decisions must be made on an empirical basis, and appropriate measures of performance capability must be employed.

The multidimensional nature of mental workload and its status as a major determinant of human performance make measures of this phenomenon essential to investigations of complex crew performance in stressful environments. However, since many of the workload questions which arise in sustained military operations must be addressed in situations which emulate these conditions as closely as possible, considerable care must be taken in the selection of a metric. Subjective ratings and behavioral secondary task measures are among the oldest and most commonly used methods of assessing workload. While each of these traditional techniques offers distinct advantages for application to a range of workload problems, neither approach is optimally suited for use in high fidelity simulation or operational test environments. The SWAT and embedded task techniques described in this paper represent attempts to refine subjective and behavioral workload assessment methodologies in order to improve the match between workload metrics and field testing requirements.

REFERENCES

(1) Navon, D. and Gopher, D. On the Economy of the Human Processing System, Psychological Review, 1979, 86(3), 214-255.

(2) Wickens, C. D. Processing Resources in Attention, Dual Task Performance, and Workload Assessment. Engineering Psychology Research Laboratory Technical Report, EPL-81-3, University of Illinois, 1981.

(3) Wierwille, W. W. and Williges, R. C. Survey and Analysis of Operator Workload Assessement Techniques. Systemetrics, Inc., S-78-101, Blacksburg, Virginia, 1978.

(4) Schiflett, S. G. Operator Workload: an Annotated Bibliography. U.S. Naval Air Test Center, SY-257R-76, Patuxent River, Maryland, 1976.

(5) Reid, G. B., Shingledecker, C. A. and Eggemeier, F. T. Application of Conjoint Measurement to Workload Scale Development. Proceedings of the 25th Annual Meeting of the Human Factors Society, Rochester, NY, 1981.

(6) Reid, G. B., Shingledecker, C. A., Nygren, T. E., and Eggemeier, F. T. Development of Multidimensional Subjective Measures of Workload. Proceedings of the International Conference on Cybernetics and Society, IEEE Systems Man and Cybernetics Society, Atlanta, Georgia, 1981.

(7) Sheridan, T. B. and Simpson, R. W. Toward the Definition and Measurement of the Mental Workload of Transport Pilots. Massachusetts Institute of Technology Flight Transportation Laboratory Report, FTL Report R79-4, Cambridge, Massachusetts, 1979.

(8) Krantz, D. H. and Tversky, A. Conjoint Measurement Analysis of Composition Rules in Psychology. Psychology Review, 1971, 78, 151-169.

(9) Shingledecker, C. A. and Crabtree, M. S. Subsidiary Radio Communications Tasks for Workload Assessment in R&D Simulations: II. Task Sensitivity Evaluation. Air Force Aerospace Medical Research Laboratory, AFAMRL-TR-82-57, Wright-Patterson Air Force Base, Ohio, 1982.

(10) Eggemeier, F. T., Crabtree, M. S., Zingg, Jennifer, Reid, G. B., and Shingledecker, C. A. Subjective Workload in a Memory Update Task. Proceedings of the 26th Annual Meeting of the Human Factors Society, Seattle, Washington, 1982.

(11) Quinn, T. J., Jauer, R. A., and Summers, P. I. Radar Aided Mission/Aircrew Capability Exploration, RAM/ACE Interim Report--Task II Synthesis. Air Force Aerospace Medical Research Laboratory, AFAMRL-TR-82-91, Wright-Patterson Air Force Base, Ohio, 1982.

(12) Broadbent, D. E. Perception and Communication, Pergamon Press, NY, 1958.

(13) Gopher, D. and North, R. A. The Measurement of Operator Capacity by Manipulation of Dual Task Demands. Aviation Research Laboratory, Report No. ARL-74-21, University of Illinois, 1974.

(14) Ogden, G. D., Levine, J. M. and Eisner, E. J. Measurement of Workload by Secondary Tasks. Human Factors, 1979, 21, 529-548.

(15) Shingledecker, C. A., Crabtree, M. S., Simons, J. C., Courtright, J. F., and O'Donnell, R. D. Subsidiary Radio Communications Tasks for Workload Assessment in R&D Simulations: I. Task Development and Scaling. Air Force Aerospace Medical Research Laboratory, AFAMRL-TR-80-126, Wright-Patterson Air Force Base, Ohio, 1980.

DISCUSSION

(Papers 5 and 6)

DR C E BILLINGS (US)

The extremely cogent presentations which we have heard provide an excellent unifying concept of workload. I am concerned, however, as to whether there is not circular reasoning with respect to task selection? I wonder if there is not an indirect approach using drugs such as hypnotics and ethanol. These substances can be employed to change performance in an extremely predictable manner and it would be possible to determine whether their effect on workload is as you would predict. Are we at the point at which pharmacological manipulation offers an indirect validation technique?

AUTHOR'S REPLY (DR C A SHINGLEDECKER (US))

No, the reasoning is not circular. Workload is an intervening variable which mediates the effects of a variety of factors on human performance. Since workload cannot be measured directly to provide a criterion for the selection of metrics, indirect methods must be used to establish validity. Most commonly, empirical validity is determined by observing the effects on a measure of several variables which we expect will impact workload in a particular fashion. This can be done by manipulation of task demands or, as you have suggested, by pharmacological manipulation. The case for validity is further bolstered when we can show that a measure also predicts performance decrement in operational behaviours which are directly related to system performance.

DR A F SANDERS (NL)

I congratulate Dr O'Donnell and Dr Shingledecker on their approach to the problem of the measurement of workload, particularly with regard to physiological measures. We have been investigating the value of the P300 of the cortical evoked response as a measure of workload in automobile driving. We found that the P300 did not correlate with difficulty of manoeuvres in driving. Does Col O'Donnell consider that the P300 is a real integrated measure of cognitive workload in a wide variety of situations? I would also like to ask Col O'Donnell and Dr Shingledecker whether you think that your subjective measures of workload will be valid for a wide variety of types of task?

AUTHOR'S REPLY (COL R D O'DONNELL (US))

I will comment on the P300 and on the physiological measures in general. Your point is very well taken that the validations of the physiological measurements have been laboratory validations. It is now time to take them out into the field. I strongly suspect that the P300 will saturate very rapidly and that it will not be appropriate for very very high stress situations in the field. However, that is why we have many measures. There are also different ways in which to use the P300. I am sold on the P300 as a laboratory measure, I still have questions as to its value in the field.

AUTHOR'S REPLY (DR C A SHINGLEDECKER (US))

Regarding the generality of the subjective technique, we have approached this area of measurement without preconceptions. We developed the general methodology and then used it in very basic laboratory tasks in which the demand was varied. We find that our subjects, which included aircrew and operators, are able to use it for various different kinds of tasks. It is certainly questionable whether it is warranted to compare SWAT scores of pilots landing a particular aircraft with the scores of subjects carrying out a simple tracking task in the laboratory. However, I do believe that subjective measures, because they provide a general perceptual impression, are probably one of the most generalised matrices available for use.

DR J SMIT (NL)

What is the real difference between the 'conventional' secondary task method and your embedded secondary task method? As far as I can see it is only a matter of acceptance by the subject.

AUTHOR'S REPLY (DR C A SHINGLEDECKER (US))

As I described them, the differences between the two methods are in their potential for primary task intrusion, instrumentation requirements, and operator acceptance. The major problem with traditional secondary tasks is that it is extremely difficult to maintain the instructed task bias that is essential to the method. The embedded task approach minimises this problem by using a secondary task which is already incorporated into the operator's concept of his work environment and which holds a lower priority in his task hierarchy than the primary task of interest.

DR C E BILLINGS (US)

With deepest apologies to John Rolfe, who was pretty much the father of the secondary task concept, I believe that we are at a point where this terminology should be abandoned. If one looks at the flying task, or indeed at any reasonably decent simulator analogue of it, one finds a considerable number of tasks whose importance varies at different points in the mission. It is possible to apply this technique to any task except that of keeping the aeroplane in the air. It is even possible that it will work with the latter task as well, particularly when it is being carried out automatically. The prioritisation of the tasks varies with the situation. Thus one is not worried about weapons configuration/weapons management in a situation in which there are no threats or one is not close to the target. Nor is one worried about communications when one is trying to stay alive because one is being chased by another aircraft carrying missiles! Thus priorities of tasks

change rapidly from moment to moment except perhaps in the sterile environment of the laboratory! I believe, therefore, that the primary, secondary, tertiary tasks approach should be abandoned.